Capstone Project Report

A Clustering Analysis on Sydney Suburbs

1. Introduction

As the largest city in Australia with a 5 million population, Sydney renowned for its diverse culture and inclusive for numerous races. Sydney naturally becomes an ideal space for start-ups who want initial their own business and a large amount of opportunities are available. The diversity and complexity of Sydney's environment thus becomes a problem that makes the people hard to decide where to establish business.

On the other hand, clustering algorithms are well-developed in recent researches and studies to derive valuable patterns from unlabelled data in a novel view. This usually contribute to discover previous unknow segmentation of data and can be used for further research in certain area such as customer segmentation can be used to formulate sensible market strategy.

In addition, many suburbs of Sydney develop distinct and diverse cultures in the long history, forming their own typical community and supporting Sydney to become one of the most liveable cities around the world. Therefore, these suburbs are the basic elements of metropolitan Sydney and can be regard as individual entities that could be used as the input to analyse the business environment of Sydney.

Subsequently, using clustering algorithms to analysis and explore the similarity and dissimilarity between Sydney suburbs could lead to discover meaningful patterns, and these patterns can provide innovative insights for the corresponding government departments or companies who are interested in developing their business in Sydney.

2. Data Collection

Data used in this study is obtained from the FourSquare Database and Australia New South Wales Government Spatial Collaboration Portal [1]. The FourSquare Database provides the detailed venue information of each suburb and the NSW Spatial Collaboration Portal offers the list of suburbs that belong to Sydney (Major part of Sydney is selected). After selection, total 400 suburbs locate in the metropolitan Sydney are included in this study, providing a comprehensive and sufficient dataset for the following clustering method.

3. Method

3.1 Data Processing

Detailed data of each suburb is obtained by saving the results of Foursquare venue query. Specifically, this query is achieved by conducting a explore search within the 500-radius circle area of the suburb centre. To successfully execute the FourSquare query the geographical coordinates of each suburb is necessary. Consequently, the suburb data downloaded from NSW government database is firstly combined with each suburb's coordinates using the Nominatim method provided by geopy library.

Then, category of venue will be extracted from the query result and frequency of different venues within the suburb will be calculated using one-hot method. Each suburb's corresponding information is grouped and transformed to one data frame. This can provide a meaningful dataset for the following K-Means Clustering.

3.2 Implement K-Means Clustering

K-Means Clustering is selected as the clustering algorithm since it is easy and fast to deploy with adequate performance that could provide an ideal clustering result.

To achieve an effective clustering performance, the Elbow method is performed to select the number of clusters which is the crucial parameter of K-Means Clustering that has a significant influence on the result. Through the evaluation, the number of clusters is decided as 3.

K-Means Clustering is performed on the input data using the scikit-learn library. The result of clustering is then be used in the following visualization process.

3.3 Visualization

To present the clustering result in a concise way that is easy for the audience to understand, the result is first visualized on the real map that display the explicit distribution of each cluster. This step is realized by merging the cluster label with original suburb data and using the folium library to create the actual map.

In addition, bar chart is selected to visualize the diversity between clusters with top five venues among each cluster. Seaborn library is used to create corresponding visualizations.

4. Results

4.1 Distribution map

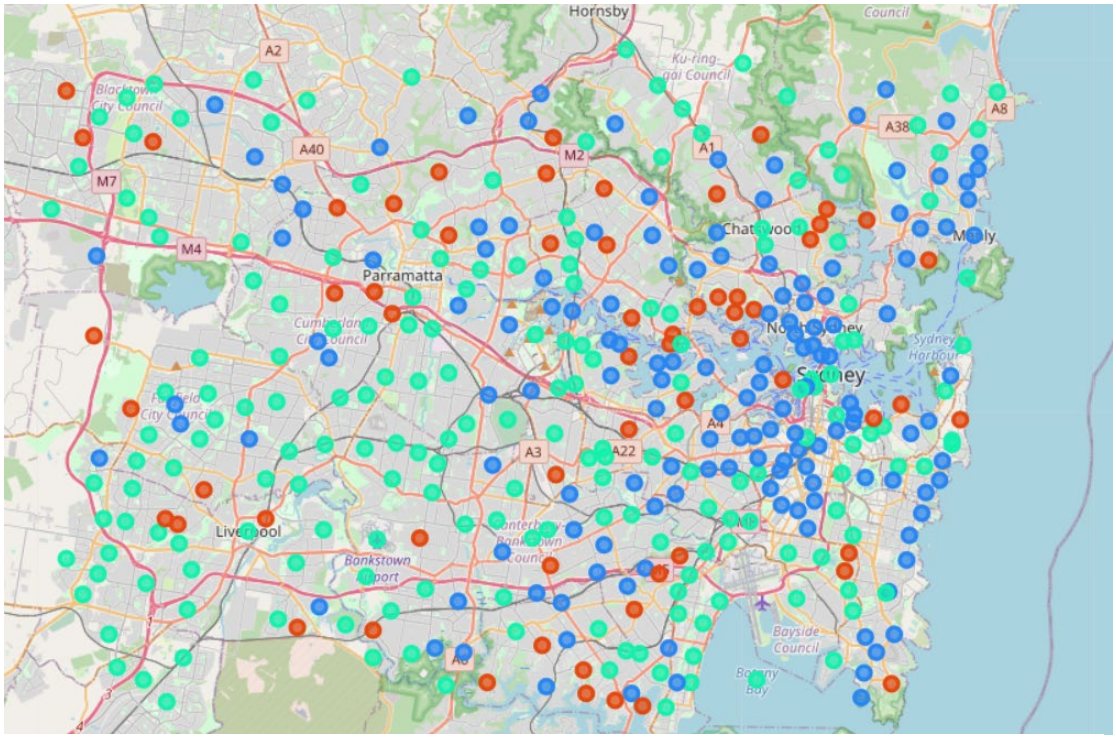

Figure 1 - Distribution map of clustering

	Number	Mark Colour
Cluster 1	146	Blue
Cluster 2	193	Fluorescent green
Cluster 3	61	Orange

The result in *Fig 1* displays the specific distribution map after the K-Means Clustering. 400 suburbs were allocated to 3 clusters with different characterizes. The largest cluster 2 has 193 suburbs, many of its suburbs are western and south of Sydney. The centre of Sydney mainly consists of suburbs that belong to cluster 1. Suburbs of cluster 3 shows a scattered distribution over the map area.

4.2 Most common venues of 3 clusters

The following bar charts on Fig 2 present the top 5 venues and their quantities among the most common venues of each cluster. Suburb groups of clusters 1 and 3 are dominated by café and parks respectively. Cluster 2 has a relatively average

distribution of different venues.

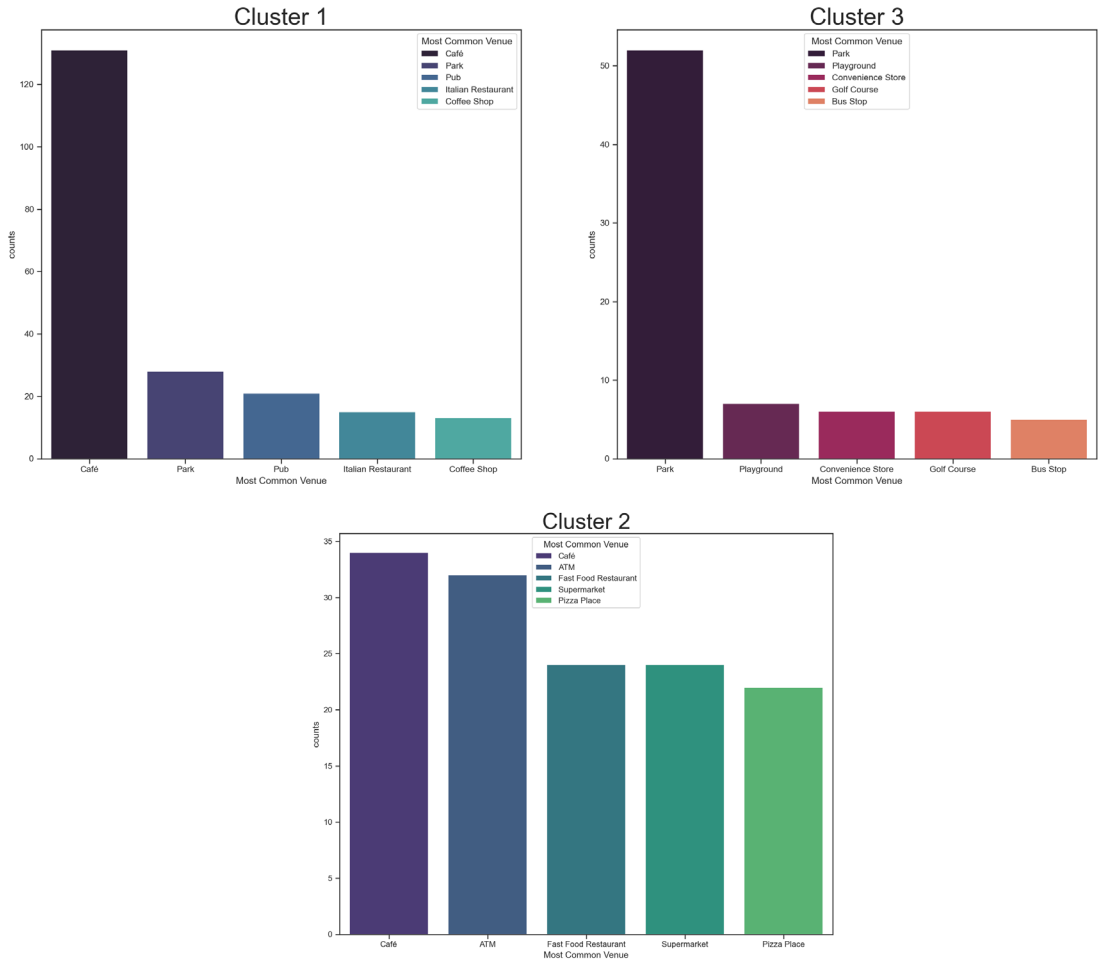

Figure 2 - Most common venues of 3 clusters

5. Discussion

Cluster 1 suburb group is governed by café which accounts for most top 5 most common venues and many of its suburbs are in the center of Sydney city area. Park and catering industry make up for the rest of the venues. This indicates that suburbs within this group might a good place for companies who are interested in foodservice. Especially for those who are planning to establish a café that serves the commuter group since it is a large group among the residents who live in the suburbs of cluster 1, these suburbs might be an ideal place to start their business.

Venues of the cluster 2 group present an average distribution among its most common venues, and these venues tend to belong to the subset of shopping centers or large shopping malls. These suburbs could be busy places where people go for their daily shopping or entertainment. This includes a lot of co-related industries such as grocery vendors. It is effective for companies who belong to these industries to focus on these suburbs to find their ideal customers.

Based on the result of the bar chart and distribution map, the suburbs of cluster 3 are suitable places for living and real estate companies since the majority of its venues are parks and playgrounds. These suburbs could have a large amount of community and population. Consequently, the government could put a special focus on these areas to enhance community security.

6. Conclusion

This study presents a detailed clustering analysis on Sydney suburbs and provides meaningful insights based on the result. There are a lot of different patterns can be discovered through the result and through this different insights or conclusions can be made.

7. Reference

1. NSW Government Spatial Collaboration Portal
https://portal.spatial.nsw.gov.au/portal/apps/sites/#/home